Straw Men and Performance Assessment

by Dr. Douglas Reeves
Advanced Learning Press, (1998)

In a recent address to the California Board of Education, Professor E.D. Hirsch offered a number of insightful comments with regard to educational reform generally and performance assessment specifically. The educational community is indebted to Professor Hirsch for his staunch advocacy of rigor and relevance in education. Unfortunately, in his recent comments before the California Board of Education, Professor Hirsch equated rigorous performance assessment with other reforms he correctly criticizes, and thus let his rhetoric outrun his logic. As an overview, I would like to note my profound respect for Professor Hirsch and the way in which he has advanced the debate over educational theory and practice. Some of the attacks on Professor Hirsch from sectors in the educational community have been ungracious, unfair, and without constructive value. With that said, I trust that I can disagree with some of Professor Hirsch's conclusions while advancing the debate and finding areas of common agreement.

In general, there is little in the evidence and examples provided by Professor Hirsch with which one can disagree. He is, in essence, against things that are silly, without rigor, and destructive. I doubt many policy-makers or educators would disagree. He has had no difficulty in finding widely publicized examples of educational practices that deserve these descriptions. Unfortunately, his conclusions about performance assessment are not warranted, any more than the existence of a few inept teachers of core knowledge programs, such as those espoused by Professor Hirsch, justify the allegation that all such programs ought to be ridiculed. The consistent error in Professor Hirsch's speech is the implication that examples of bad performance assessment represent all such practices, and examples of good multiple choice assessment accurately represent all such tests.

In the following five points, I outline three areas of disagreement and two areas of agreement with Professor Hirsch's address :

- 1 -

Generalizations About the Reliability of Multiple Choice Tests and the Unreliability of Performance Assessments Are Misleading and Inaccurate

Professor Hirsch, as with many critics of performance assessments, appeals to the statistical "reliability" – that is, consistency of evaluation and responses – of tests as an appropriate criteria for evaluating the usefulness of tests. I agree with this premise, but not the conclusions that stem from it. Multiple choice assessments are not inherently reliable and performance assessments are not inherently unreliable. The mere existence of a limited number of correct answers, as multiple choice tests provide, do not guarantee statistical reliability. That is the heart of the research by Dr. Lee Cronbach of Stanford University, to whom Professor Hirsch refers with approval. Multiple choice tests can have low reliability when their directions are misunderstood, when items in the same test are interpreted differently by students, and when the number of students taking the test is small. Note well: Even multiple choice tests that boast high reliability coefficients (often measured by "Cronbach's alpha") are most often based on administration to large numbers of students. Whether the same degree of consistency would be measured at the school or classroom level is a frequently unanswered question. There are, to be sure, some dreadful performance assessments that yield inconsistent – that is, statistically unreliable – results, just as there are performance assessments that yield consistent and statistically reliable results. As with multiple choice tests, high reliability is a function of specificity, careful development, and good administration of the test – it is not a function of the form of the test. Professor Hirsch's evidence

makes a case for good performance assessments and good multiple choice tests and for avoiding bad tests of any sort – his evidence, and the research literature to which he refers, does not support wholesale categorizations of tests based on their format.

Professor Hirsch's emphasis on reliability is only a starting point for the criteria for effective testing programs. An issue of equal significance to that of reliability is validity – does the test measure what we think it does? School children and teachers spend countless hours taking – and parents and policymakers spend more hours analyzing – tests that are reliable but not valid. I can give a test in Latvian to prospective firefighters and get statistically consistent results – no one would suggest they are valid. The educational equivalent of this happens when school systems demand, on the one hand, broadly based curriculums with high sounding standards for communication skills along with the ability of students to explain answers. These same school systems demand, on the other hand, tests that measure a tiny fraction of those skills, or do not measure them at all. They get statistically consistent – reliable – results, and make multi-million dollar policy decisions based on tests that do not test what they think they are testing. This is an example of the ill-advised experiments that Professor Hirsch correctly condemns, but appears nevertheless to recommend in his speech.

In his praise of statistical properties of effective tests, Professor Hirsch appears to put advocates of performance assessment and psycho-metricians on opposite ends of the continuum. In fact, responsible advocates of performance assessment insist on rigorous statistical measurements of reliability and embrace the psychometric developments of Dr. Cronbach, Bob Linn, Eva Baker, and the others to whom Professor Hirsch refers. Yes, there are performance assessments that are unreliable and invalid, advocated by a legion of consultants more concerned with politics than accuracy. But the existence of charlatans in any field does not justify their use as straw men to make unwarranted generalizations.

Finally, the nature of measuring reliability assumes wide variation among students. Consider the example of a teacher who administered a challenging and relevant test after preparing the class well and after the students have studied diligently. If all or almost all of the students respond correctly to the questions on the test, I am quite confident of two things. First, Cronbach's alpha – the reliability coefficient-will be at or near zero. Second, the teacher – perhaps even Lee Cronbach or E.D. Hirsch – would deserve congratulations, not condemnation for having an "unreliable" test. Basing an argument only on statistical reliability demanding wide variation among students is an inappropriate approach, particularly in a state such as Virginia where, as a matter of policy, academic standards have been formulated with the explicit expectation that all students should meet them.

Performance Assessments and Multiple Choice Tests are Not Mutually Exclusive

There are those who make one side or the other in this debate appear to be the Darth Vadar's of education, representing the dark side of the force with their peculiar tests. In fact, the world of today's students and workers requires both types of tests. Regardless of one's predisposition in favor of one type of testing or another, the multiple choice format will remain predominant for admission to college, graduate, and professional school. Performance tests will remain predominant in many college classes, technical trades, and professions. Responsible educators should demand both. The suggestion that "performance assessment is acceptable for classroom use" along with the demand that different tests should be used for graduation and "real" assessment is the essence of invalidity. We must teach what our policymakers require – a variety of skills, including core knowledge, analysis, and communication – and we must test in ways that accurately assess those skills. The "either/or" construction of Professor Hirsch's arguments creates

a dichotomy where one need not exist.

- 3 -

The Indictment of Math Performance Assessment is Based on the False Assumption that Performance Assessment Excludes Accuracy and Rigor

Professor Hirsch appears to assume that the appropriate imperative for mathematical accuracy is at odds with the need for conceptual understanding of mathematics. The existence of highly publicized bad teaching in this area does not justify the conclusions drawn by Professor Hirsch. Responsible advocates of performance assessment do not regard accuracy in number operations as unessential. Indeed, any math performance assessment I have ever administered required accuracy – not merely "concepts and methods" which are apparently maligned by Professor Hirsch. As a teacher and professor, I wanted my math students – from grade school to graduate school – to explain their responses and apply the mathematical knowledge they had to new and unfamiliar problems. Moreover, I expected more of the teacher – not merely to "check" the answer, but to diagnose the student response to determine whether the students' errors were in number operations, conceptual understanding, or application. I would submit that my approach to math education is the one requiring more rigor on the part of both teacher and student.

- 4 -

Professor Hirsch is Correct in Endorsing the Use of Challenging Educational Practices, Rather than "Developmentally Appropriate" Material

I strongly agree with Professor Hirsch in his condemnation of the use of insufficiently rigorous and challenging content for students of any age. Indeed, studies dating back to the 1960s justify the use of high challenges and high expectations by teachers. One of the primary justifications for performance assessment is that it is more challenging and more rigorous than multiple choice tests, because performance tasks demand explanations, analysis, and application that multiple choice tests fail to provide. The existence of feeble performance assessments, along with unchallenging and silly multiple choice tests, does not justify the abandonment of either form. We both agree on the fundamental need for rigor.

- 5 -

Professor Hirsch is Correct in His Critique of Educational Research, Though Research Can be Carefully Screened and Instructive for Educational Policy

Professor Hirsch and I would agree on the faulty nature of much of what passes for "research" in the educational community. One must wade through a lot of chaff to find the few grains of wheat. A recent (1997) review by Ellis and Fouts entitled Research on Educational Innovations would perhaps be useful. Along with Professor Hirsch, the reviewers find fault with many educational fads, including some poorly constructed notions of self-esteem. But in the objective fashion we should all favor, they find much to commend in the 20-year research base accumulated at Johns Hopkins University and elsewhere in support of cooperative learning. Unfortunately, many critics of educational reform throw the baby out with the bath water, rejecting those reforms that are demonstrably effective – and cooperative learning and performance assessment are in that category. It is fair to note that a growing body of research supports the use of explicit academic curriculum standards, such as the core knowledge programs of Professor Hirsch. These programs do not exclude performance assessment, and the several core knowledge schools I have observed take the balanced approach to assessment – requiring both multiple choice tests and performance assessments.

Perhaps the most important point from a state policy standpoint is this: Test scores alone cannot be the indicator of whether effective educational practices are in place. Professor Hirsch is an

eloquent advocate of rigorous teaching practices, curriculum content, and student requirements. An effective accountability system would include indications of the extent to which these demonstrably effective strategies are in use. Two schools with similar test scores could be viewed through much different light if an accountability system also allowed policy makers and educational leaders to evaluate the extent to which teachers and administrators were undertaking steps to make improvements.

Finally, consider the interaction between testing and standards. Virginia leads the nation in the articulation of educational standards. Although there is always room for improvement, the clarity of language, high rigor, and uniform expectations of all students contained in the Virginia standards is a model for other states to follow. These standards will be rendered another set of empty slogans if the state adopts a testing and accountability system that is inconsistent with the standards. Economic and practical considerations are real – but the stakes are high. The testing system adopted by the state is itself a test – a test of the integrity of the standards. A comprehensive accountability and assessment system, including both multiple choice and performance assessments, can be constructed within the bounds of economy and practicality. I would hope that Professor Hirsch, as a leading advocate of educational rigor, would be an architect of such a system, rather than opponent of it.